Distr. GENERAL
02 September 2013
WP 30
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on Statistical Data Collection
(Geneva, Switzerland, 25-27 September 2013)

Topic (v): Integration and management of new data sources

BIG DATA FOR OFFICIAL STATISTICS – STRATEGIES AND SOME INITIAL EUROPEAN APPLICATIONS

Working Paper

Prepared by Martin Karlberg and Michail Skaliotis, Eurostat[1]

"We need to promote the *greater statistics* brand" – David J. Hand, 2008

I. Introduction

1. The official statistics profession is rather conservative and very cautious about engaging itself in novel types of data and new fields. Big Data is no exemption. For several years developments in Big Data were largely driven by computer scientists. This is still the case today. One of the consequences of this conservatism and extreme cautiousness is what Professor D.J. Hand described in the following: *Indeed, it is a great pity that the snappy phrase "information technology" has been appropriated by the computer scientists. The truth is that computer scientists are really data technologists, concerned with storing and manipulating data. But it is statisticians for whom the raison d'être is the extraction of meaning from data-whose job is to transform data into information.*

2. Can we say that Big Data represent a *missed opportunity* for official statisticians? We do not think so. While there has been a serious delay in getting statistics agencies on board, recent evidence suggests that the (lost) ground can be regained soon. Several National Statistical Institutes (NSIs) and leading international bodies including UNSD, UNECE, OECD and Eurostat have undertaken important initiatives which manifest that Big Data is high on their agenda. For instance, two initiatives led by the UNECE in this regard merit mention: (i) a strategic paper[2] which identifies the main challenges with regard to legislation, privacy, financial issues, management, methodology and technology, and provides some basic recommendations for national and international statistical bodies; (ii) a follow-up aiming at formulating a concrete project proposal to effectively address some of the challenges identified in the strategic paper. The project proposal[3] has three main objectives: (a) *to identify, examine and provide guidance for statistical organizations to act upon the main strategic and*

[1] The authors wish to extend their particular thanks to Albrecht Wirthmann for his in-depth contributions concerning the price collection project. Any errors and omissions are the sole responsibility of the authors; the opinions expressed in this paper are personal and do not necessarily reflect the official position of the European Commission.
[2] *What does Big Data mean for Official Statistics*; published 10 March 2013, and available on http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622
[3] *The Role of Big Data in the Modernisation of Statistical Production*, jointly elaborated by 13 national and international and international statistics organisations; currently in its final drafting stage; available on http://www1.unece.org/stat/platform/display/Collection/Draft+HLG+Project+Proposal+on+Big+Data

methodological issues that Big Data poses for the official statistics industry (b) to demonstrate the feasibility of efficient production of both novel products and "mainstream" official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts, and (c) to facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.

3. Eurostat's activities – which are the focus of this paper – also illustrate the significant change in the attitude of official statisticians towards Big Data. In addition to the projects (see II below), it is important to underline that this year's DGINS[4] conference[5] is devoted to the theme of Big Data and Official Statistics. The Heads of NSIs of the European Statistical System (ESS) will discuss a common strategy towards Big Data and it is foreseen that they will adopt a declaration (*The Scheveningen Memorandum*) which would constitute a basis for joint actions in the years to come.

4. The overarching objective of this paper is to make a small contribution in the debate around the question of *"How should Official Statistics evolve in the Age of Big Data in order to remain relevant?"* We do not claim having the magic answers, but we are very much convinced that this new evolution in official statistics will be driven by concrete applications in a variety of fields. Therefore, we focus on three projects launched by Eurostat in 2013. As could be seen in II below, even if most of these projects are in an early phase, fundamental issues regarding privacy, methodology, trust, access to third party data, skills development, technology issues etc. have already emerged.

5. We also believe that time has come to demystify Big Data and embrace it as part of our business. This may initially require a paradigm shift in many respects (novel data sources and statistical approaches), but essentially we shall continue to deal with "extracting information/value from data". The fact that we will deal with different types of data, data of enormous volumes, dynamically and continuously streamed, automatically captured, data of unknown structures, should not be seen as barriers but rather as methodological challenges and opportunities for our profession and the science of statistics in general. Following Chambers' (1993)[6] definition of "greater statistics", and Hand's (2008)[7] strong suggestions and convincing arguments about applying this definition into the current dynamic data ecosystem, we would like concur in stating that Big Data should be considered as being an integral part of *greater statistics*, part of our everyday business. Chambers defines greater statistics as *"everything related to learning from data, from the first planning or collection to the presentation or report"*. Hand is arguing, following Chamber's greater statistics, that areas like *neural nets, rule-based methods, tree-based algorithms, genetic algorithms, fuzzy logic, mixture models, Bayesian networks, and meta-learning,* should be regarded as part of statistics. We agree, while adding that developments in Big Data analytics should be regarded as part of (greater) statistics.

II. Initial European Applications

A. Using the Internet for the collection of information society and other statistics

1. Drivers for improvement of Information society statistics

6. Eurostat compiles and disseminates a wide variety of Information and Communication Technology (ICT) statistics, which are used to monitor the progression of European countries to Information Societies. Currently, the underlying data are compiled by means of two different questionnaires (one for households/individuals and one for enterprises), both of which are administered by means of traditional surveys of people and businesses.

7. In all policy areas covered by official statistics, user needs evolve as society evolves. However, an issue of particular relevance to the ICT statistics is the fact that the phenomena under study evolve comparatively rapidly. For instance, the increased penetration of e.g. computers and broadband

[4] The acronym DGINS originates from French and denotes the "General Directors of National Statistical Institutes"
[5] http://www.cbs-events.nl/dgins2013/dgins-2013-information/
[6] Chambers, J.M. (1993) Greater or lesser statistics: a choice for future research. *Statist. Comput., 3, 182-184.*
[7] Hand, D.J. Modern statistics: the myth and the magic. *Journal of the .R. Stat. Society A*(2009), 172/2, pp. 287-306.

connections render certain of the indicators on to access to ICT (and to the Internet) less relevant. On the other hand, new technology constantly becomes available, and therefore, questions on access and use of these new technologies of become relevant to policymakers.

8. Digital footprints are left behind in our daily lives, and can be used to measure a wide variety of phenomena. A recently conducted study by the European Commission's Directorate-General for Communication Networks, Content and Technology (DG Connect) investigated the possibility to use the Internet as a Data Source (IaD) to complement or substitute traditional statistical sources.

9. Three basic types of IaD methods are identified in the IaD model: (i) User-centric measurements that capture changes in behaviour at the client side (PC, smartphone) of an individual user; (ii) Network-centric measurements that focus on measuring properties of the underlying network; (iii) Site-centric measurements that obtain data from webservers.

10. Any official statistics related to phenomena generating a digital footprint could be eligible for reengineering to make use of such a footprint. However, it could be argued that by definition, digital footprints all have been generated through the active or passive use of ICT, and therefore would lend themselves particularly well to use for the purpose of compiling statistics on ICT usage.

11. Given the rapidly evolving needs (7) on one hand and the evolution of new potential sources at a similarly rapid rate (8, 10) on the other hand, ICT statistics is a natural candidate subject for piloting reengineering based on Internet and similar sources. Eurostat has therefore commissioned a study on the analysis of methodologies for using the Internet for the collection of ICT and other statistics.

2. Scope

12. An important aspect of the analysis is to identify an effective set of transformation and validation rules that will render data of the IaD variety of a high quality (timeliness/punctuality, accuracy, coherence/comparability). Based on the past experience in developing Internet and Web standards, the study proposes refraining from all-encompassing transformation and validation rules, and favours an incremental approach, subdividing the general problem into minor issues (such as IaD for specific ICT statistics). Moreover, a "procrastination" approach is advocated, with only minimum necessary development work taking place early since stakeholders may self-organize and drive development.

13. Out of the three possible IaD methods (9), the methods proposed in the study could be said to fall in the user-centric and site-centric categories.

14. The study uses a wide definition of "Internet as a data source", since with the exception of data on enterprise websites, much of the data are not necessarily stored on the Internet, but rather related to Internet use – and possible to (once compiled) transfer over the Internet.

3. Feasibility

15. The study has gone over virtually all items of the current household/individual and enterprise surveys, and assessed their suitability for collection over the Internet. A first conclusion is that items trying to elicit opinions (e.g. rationale for use/non-use, level of satisfaction) and "offline" items (e.g. non-ICT use of government services) are not possible to measure based on digital footprints.

16. For individuals/households it is concluded that neither access to ICT (module A) nor use of computers (module B) is possible to measure over the Internet, whereas use of Internet (module C) quite predictably is ideal for measurement over the Internet. For use of e-government (module D), measurement over the Internet is only possible after quite some redefinition of the items. For e-commerce (module E) and e-skills (module F), collection over the Internet is deemed to be feasible.

17. For enterprises, it is concluded that only a fraction of the ICT variables would be available on the enterprise website, and that most information, albeit stored or possible to trace by means if ICT tools, would only be available via the servers of the enterprise.

4. New potential ICT indicators

18. The study concludes that the multipurpose nature of data collected electronically would provide ample material for additional ICT indicators, defined ex-post, and a large number of potential indicators, providing additional detail to existing items are suggested (e.g. "number of e-mails sent" under "types of use").

5. Proposed implementation

(a) Proposed implementation for households/individuals

19. As a number of indicators would have to be redefined (and as the collection mode changes) to allow them to be collected over the Internet, a time series break is inevitable. The study therefore suggests administering the current and future surveys simultaneously at least once, at least for a few "representative" countries, to allow for "bridging" the gap between the two series.

20. For households/individuals, the study proposes a rather traditional approach, with respondents contacted via telephone or mail according to current country practice. Items not possible to collect with an IaD approach are administered in the same way as before. Using access to the internet as a "portal" question, the respondents having Internet access are invited to download a program (application) which would monitor their Internet activities. In theory, this tool could be used for "offline" activities as well, thus rendering module B (use of computers) possible to measure as well.

21. The monitoring tool would transmit the stored data at a pre-set frequency (e.g. real-time, daily, at the end of the reference period) to a server at the NSI in question. For privacy reasons, the respondents would have the possibility to turn of the monitoring tool at any time. Incomplete data resulting from this would have to be analysed through the traditional methods available for item non-response.

(b) Proposed implementation for enterprises

22. The study proposes a phased implementation for enterprises. In the first step, which focuses on harvesting data from enterprise websites, automated searches for websites are conducted, and URLs thus identified are entered and maintained in the national Business Registers. Some subsequent quality control would complement this automatic procedure to reduce the number of errors.

23. In the second phase, for enterprises with websites, data would either (following consent) be scraped or harvested by means of a crawler – or obtained by means of a program provided to the enterprise.

24. In a possible third phase, the server data (needed for the majority of the ICT indicators of the enterprise) would be collected – either by means of a program/application, or by the provision of server log files from the enterprise to the NSI.

B. Use of mobile positioning data for tourism statistics

1. Drivers for basing tourism statistics on mobile positioning data

25. There is a high interest in basing tourism statistics on mobile positioning-based data sources. In a stock-taking undertaken within the *Feasibility study on the use of mobile positioning data for tourism statistics* (commissioned by Eurostat), the following reasons are quoted: (i) a more accurate and in-depth data source for describing the behaviour of people within a destination (in tourism research); (ii) new aspects and indicators for describing the time-space behaviour of people; (iii) a highly quantitative data source with many options and methods to process and analyse; (iv) better time-space insights compared to traditional methods; (v) cheaper compared to other data collection methods at the same scale; (vi) possibility to register trips to sparsely populated locations such as natural parks, for which it is difficult to reach visitors with questionnaires or counters.

26. On the other hand, if mobile positioning statistics are used, there are issues with qualitative information (on the user, on the purpose of the visits, on the means of transport).

2. Current use

27. The number of studies, research papers, projects and applications based on mobile data is increasing. A number of NSIs have started exploring opportunities for using such data for tourism statistics.

28. The Central Bank of Estonia has been using state-level inbound and outbound tourism statistics (trips, spent nights) based on mobile positioning data and calibrated with official accommodation and travel statistics since 2009, when a feasibility study on using the data in Estonia was conducted. The monthly data flow is used in the calculation of the national balance of payments. The initiation of this need came from border surveys being cancelled due to financial cutbacks.

3. Data access – the main barrier

29. Most studies conducted to date have been experimental, with varying arrangements for access to mobile positioning data. Now, if mobile positioning data are to become a viable option for producing comparable European tourism statistics, the data access issue must be tackled in a concerted way, and the business continuity of tourism statistics – and thus the continued provision of mobile positioning data to NSIs – has to be guaranteed. To this end, the feasibility study contains a specific work package on the feasibility of access. The conclusions reached so far are summarised below.

30. Privacy concerns: although there has been a cultural shift, with people being increasingly willing to share, or even actively disseminate (Facebook) their personal data, large-scale provision of mobile positioning data to government agencies could be perceived as an invasion of privacy.

31. Data protection legislation: there are a number of EU-level instruments with different aims (data protection vs. data retention). Moreover, the production of official statistics could possibly be considered as a statutory basis for which processing of personal data (such as mobile positioning data) could take place. To further complicate matters, different European countries have different national transpositions of the EU directives.

32. Data provider reluctance: while mobile network operators (MNOs) have an interest in this initiative, there are issues concerning the (i) the maintenance of business secrets; (ii) direct costs (in terms of human and financial resources) of providing data (iii) effect of the data extraction workload on the real-time systems; (iv) the opportunity cost of giving away their data "for free" (possibly by means of binding legislation) to NSIs.

33. Technological barriers: as there are a number of MNOs in each country, a technical solution concerning how to merge data from different MNOs into one single analysis data set need to be resolved. This particularly concerns at which step anonymisation should take place (anonymisation far upstream maximises privacy, but complicates processing and leads to loss of information).

4. Technological and methodological issues

34. Leaving the legal and ethical aspects aside, there are a number of other, less controversial issues which need to be addressed. Issues identified in the access feasibility work package are listed below.

35. Data standardisation: Technical platforms and data formats may differ between MNOs, but the data provided by must be standardised – both in terms of format, content and frequency (temporal granularity). The format, content and attributes of the data should be stable over time.

36. Provision frequency: The frequency by which the MNOs transmit data to NSIs (near-real time, daily, weekly, monthly etc.) should be defined. Scalability and speed of the processing is also an issue – the processing time should be independent of the size of the operator (with an increase in parallel processing resources compensating for operator size).

37. Methodology and quality: Algorithms should be the same across MNOs, the issue of sampling and representativity needs to be tackled, possibly by means of calibration as is done in Estonia (28). Quality assurance (including corrective action) should be conducted at most process steps.

C. Collecting prices on the internet

1. Background

38. Official statistics have a long tradition of collecting price information for goods and services. Prices and their interrelationships make up the system of prices, which affect all sections of the society and economy and determine the allocation of resources and consumption patterns. The system of price statistics satisfies the needs of a variety of users and is essential for monitoring the development of the economy, defining monetary policy, deciding on investments and expenditures.

39. The consumer price index is the most popular statistical indicator. It provides a measure for the development of prices from the consumer's perspective and is derived from prices of products typically purchased by a household.

40. The fast development of the internet during the last decade has led to the widespread adoption of ecommerce by producers and consumers. Nowadays, consumers are able to buy a wide variety of goods and services from webshops at the internet. The share of European citizens who ordered goods or services online has doubled from 15% in 2005 to 35% in 2012[8]. Online purchasing is getting a common activity and is supplementing or even replacing purchases via physical shops.

41. A consequence of this development is that almost all products that are bought by a household are offered via webshops on the internet. Product information is usually given on the type of product and its price together with a short description. This allows for using the internet for price collection.

2. Projects at national and European level

42. Some NSIs within the European Union have conducted feasibility studies on collecting prices from internet webshops with the Dutch statistical office being the most advanced. Eurostat has reacted on these initiatives and called for projects at European level. The aim of the initiative is to promote feasibility studies regarding the automated collection of detailed prices data from the internet across European countries. The studies should assess methodological problems/issues related to the internet collection of detailed prices, assess results of data collections and outline areas for further research.

43. Statistics Netherlands (CBS) is conducting a project for collecting prices for airline tickets, cloths, real estate and cinema tickets from the internet. The project started in January 2013 and will last for two years. Within the project, CBS will develop generic software modules for internet prices collection under open source licence. In this context, "generic" means that the software would be able to collect prices of different products in different countries.

44. Prices are collected via internet robots that visit websites, analyse their contents, search for the relevant information and return, if successful, the requested information. This approach benefits from the fact that many webshops have the same structure. A user can search for products, gets a list of results with a short description of each item including its price or can click for more detail.

45. The main drivers for this initiative are the intention to reduce costs of data collection, reduce burden on respondents and to produce results much faster as compared to traditional prices collections.

3. Issues

46. Problems are arising from the dynamics of websites and items, different categorisation or limited comparability due to lack of standardisation. These issues contribute to higher complexity in programming generic software for price collection via the internet.

47. Further methodological issues might be differences between online and offline prices and the volatility of online prices. Price differences can be assessed by comparing online and offline prices.

[8] Eurostat Statistics on ICT usage in households, http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&tableSelection=1&labeling=labels&footnotes=yes&layout=time,geo,cat&language=en&pcode=tin00067&plugin=1

First results are promising. With internet robots, it is possible to retrieve prices very frequently exceeding by far the frequency of offline price collection. Analyses of online prices reveal the dynamics of prices which has to be treated to derive stable and robust statistic.

48. The studies conducted so far show clearly the potential of internet price collection for the purpose of deriving price indices. The suitability of the method depends on the type of product. Given the current expansion of e-commerce this method could surely be extended to other product groups.

D. Reflections on the European applications so far

1. Are these Big Data applications?

49. While the Tourism project deals with data which are rather structured in nature, it is also clearly dealing with transaction data which in principle already are present and are being stored, irrespective of the production of official statistics. It thus comes close to the notion of "organic data". Moreover, the volume of transaction data is huge, and it could credibly be argued that these are indeed Big Data.

50. In contrast, the IaD study deals with a large variety of data. With the exception of the enterprise website data, much of the data sources do not exist currently – measurement would only commence once respondents installs a monitoring tool/application on their devices. In spite of the unstructured nature of the data generated through automatic monitoring, this thus comes closer to the notion of "designed data". (However, the upcoming stages of the IaD project include a strand on Federated Open Data, which would clearly fall in the field of Big Data.)

51. As Internet Price Collection initiatives clearly deal with large volumes of heterogeneous data made available to the general public independent of official statistics, the Big Data term clearly applies.

2. Representativity issues

(a) One company – one website?

52. The IaD study may underestimate the complexity of the issue of identifying the websites of enterprises, and entering those in the national Business Registers. While certain SMEs may indeed have a single website for all their business, larger companies could have a multitude of websites, not the least those with an intensive interaction with consumers, those with a large multinational presence, and those with a wide range of brands in their portfolio. The correlation between ICT usage and the complexity of web presence further aggravates this issue.

53. Conducting an appropriate inventory of the Web presence of such enterprises may prove a gargantuan task. Some parallels could perhaps be drawn to the ESSnet on Profiling[9].

54. As an alternative (or complement) to web searches, domain name registry data might perhaps be used to obtain information on websites registered by enterprises.

(b) One person – one computer?

55. Appropriate consideration has to be taken of the fact that nowadays, ICT usage may take place via a number of devices. "Monitoring software" would thus have to be installed on all devices "in scope".

(c) Alternative sources?

56. To gauge e-commerce volumes, the possible last (server data/log data; 24) phase of enterprise data collection may prove a rather blunt tool, so the alternative of basing such ICT indicators on direct export of data from the business/accounting system of the enterprise (for enterprises having appropriate systems) could be investigated.

[9] http://www.cros-portal.eu/content/profiling

(d) Cell phone penetration

57. While the general issue of representativity is raised in the Tourism study, the cell phone penetration rate is not explicitly discussed. However, a reasonable conjecture is that the cell phone penetration in the countries under study is so high, that "not much is missed out" in relation to traditional surveys.

3. Other quality issues

(a) Comparability vs. relevance

58. With new methods, a time series break is inevitable, and bridging based on parallel measurement is proposed (19; 47) to allow for some comparability over time. However, once the one-off investment in collection of multipurpose data has been made, the new statistics may (depending on the granularity of the data collected automatically) become more responsive to user needs, and reduce the lead times for developing new indicators. As stated in the IaD study, "Contemplating departures from the existing way of compiling and dissemination indictors into the new ways made possible by the digital footprints (in particular the Internet) is a necessary bridge to the future."

(b) Coherence

59. In the IaD study, coherence is declared a non-issue, on the grounds that the current ICT indicators are of a standalone nature. However, given the multipurpose nature of the data collected in the IaD project, new indicators, consistent by definition (since they would be based on a single dataset) might emerge, thereby improving coherence of ICT statistics.

4. Data provision issues

(a) Ethical and image issues

60. For the IaD project, it might be hard to gain acceptance among individuals for the collection mode (as indicated in I.A.5(a), this concerns installation of monitoring software), due to its very strong resemblance to phishing. Public-awareness campaigns have made people reluctant (and rightly so) to install what could be perceived as "spyware" on their devices.

61. The possible last (server data/log data; 24) phase of enterprise data collection may, depending on the granularity of the data collection, actually generate data on employee level. Thereby, the data collected could no longer be considered purely enterprise data, but rather personal data – so regulations concerning the protection of personal data come into play. This needs further study.

62. In terms of business secrets, the Tourism study notes that MNOs are concerned about the possibility that competitors acquiring information concerning e.g. the number of subscribers and the number of service activities (calls, messaging, data), and foresees strict procedures to be put in place so that such information does not leak to competitors. However, it might prove very hard to prevent the protection of such information because of the oligopoly situation in many countries, with only a handful of operators. This thus becomes a statistical disclosure control issue, with cells having very low (3-5?) frequencies. It could be argued that if the algorithms for generating statistics are reasonably transparent, MNOs may reverse-engineer national tourism statistics, based on their own data, and thus extrapolate the underlying data for the other (2-4?) major competitors in that country. It should be demonstrated how such disclosure risks could be eliminated.

63. For the internet price collection initiatives, legal issues arise from using internet robots for harvesting information from websites. The approach of the statistical offices would be to be as transparent as possible, for instance informing webshops on the fact that their websites are being harvested. However, information might prove insufficient, since some webshops explicitly forbid robot use.

(b) Technical issues

64. The Tourism project has identified "effect of the data extraction workload on the real-time systems" as an issue to be tackled. This is equally true for the IaD project – and perhaps even a higher risk. For Tourism, a handful of professional data providers (MNOs), with a limited number of systems, is concerned, whereas a large number of data providers (private individuals), with a very large number of different system configurations, are concerned. There is a clear risk that monitoring tools "freeze up" the systems of private individuals (or use up free storage space etc.)

65. If it proves to be a real issue, the effect of the extraction of mobile positioning data on real-time systems could be mitigated by means of sampling – in principle, given the magnitude of the data collected, it seems reasonable to assume that a subset (possibly oversampling smaller domains) would still suffice to generate accurate estimates, as long as the sampling procedure is well-defined.

(c) Continuity issues

66. As stated above (29), if mobile positioning data are to become a viable option for producing comparable European tourism statistics, the continued provision of mobile positioning data to NSIs must be guaranteed. It is hard to see any alternatives to legislation if all MNOs should provide data.

67. For IaD-based statistics, continuity might be even harder to ascertain. The fragmented and heterogeneous data harvested would have to be continually cleaned, and the acceptance for "monitoring tools" might be hard to achieve. Hopefully, the general trend (30) towards increased sharing (and the fact that having a "monitoring tool" could be perceived as less burdensome than answering questions) would facilitate acceptance.

68. For price collection via the internet, the software has to be adapted to the changing content and structure of the internet. New webshops might appear while others disappear. Structure of websites might change and the contents of the websites are also subject to change.

5. Learning by doing

69. Above, a number of issues concerning the three studies have been raised. However, while this might give the impression that doing nothing might be a better bet than embarking on a journey with an uncertain outcome, the cost of doing nothing shouldn't be ignored; continuing to compile statistics by means of traditional surveys and ignoring emerging sources might render certain official statistics obsolete and might lead to many lost opportunities.

70. Desktop analyses can only bring us so far; we also need to actually try out different approaches to see what works in practice and which improvements are needed. During the course of the projects, not the least the pilot study planned in the IaD project, new knowledge will be gathered, and the practical feasibility of the approaches under development will be assessed.

III. A Big Data Strategy for Official Statistics? Issues and Conclusions

71. It is almost certain that the current *momentum* of increased national and international interest in Big Data and statistics will inevitably encourage NSIs (and international statistical agencies) to adopt dedicated programmes and strategies in this field. Planning a Big Data strategy for official statistics is not business as usual, i.e. it is not a traditional multi-annual strategic programme. There are many aspects which should be carefully considered in such an operation; a detailed enumeration and analysis of the key issues would require – on its own – much more space than the current document. We have therefore decided to select six issues which we believe are fundamental in designing a strategic action programme in this area.

72. *A strategy for Big Data is much wider than Official Statistics*: We are convinced that a Big Data strategy for official statistics should be an integral part of a national (or international)[10] strategy. Big Data has attracted the interest of almost all policy areas that see enormous opportunities for better service delivery, policy development and monitoring, enhanced accountability and productivity gains. Official statistics is one amongst many policy fields; while we should focus on our analytics and statistical methodology strengths it is equally important to be an active member of a wider public sector Big Data strategy. Many of the challenges and issues described above within the context of Eurostat's projects (technology, privacy, legislation, etc.) are common to many fields and therefore can be better addressed at a national/government coordinated level.

73. *Develop strategic partnerships with industry and academia (Public-Private-Partnerships; PPPs)*: Collaboration with owners of private data sources, technology leaders, and academic / research institutions on a win-win basis is absolutely essential in many respects. Business and academics have been working on data analytics developments for some time and we can leverage this experience and expertise while adding our own strengths regarding statistical methodology, quality, handling confidential data, etc. A small number of strategic alliances in the form of PPPs will also help to build trust amongst official statistical agencies and owners of third party data sets – a playground of increased importance for all those dealing with Big Data analytics.

74. *Demystify "data science" – develop the necessary skills and internal analytic capabilities*: Rather than rehashing the issue of serious shortages of data scientists in the near future, we wish to encourage NSIs to enhance the skills of their staff members through targeted training. Our own experience in Eurostat with a small number of staff following some basic Big Data analytics courses is very positive. Considering the mix of skills which are required for "data scientists", we conclude that a statistician with a working knowledge of modern programming languages could qualify much easier than an IT specialist who lacks training in statistics. Moreover, most NSIs already have staff members with advanced IT skills who could assure the effective transfer of knowledge.

75. *Adopt an "applications-driven" approach*: We have emphasised throughout this paper that the best way for official statistics to stay relevant in the era of Big Data is through applications and projects in various policy areas (see I.D.5). In designing and preparing such projects, NSIs should leverage the good collaboration and co-ordination with other public sector, industry and academic bodies (see 73 and 74). A particular field of interest for governments in this regard is "open data and PSI", i.e. using Big Data analytics within the enormous data.gov and public sector information (PSI) repositories.

76. *Adopt Privacy by design and basic principles of responsible analytics*: We should seek advice by legal experts (include them as members of the project teams) in order to ensure that privacy protection and data confidentiality is properly handled throughout the life cycle of any Big Data project. NSIs can leverage their long experience and public trust in dealing with confidential data and should play a central role within a national Big Data strategy.

77. *Promote the concept of greater statistics*: We underlined earlier (5) how important it is for statisticians to make this "mental shift" and embrace Big Data as being part of their business, part of what Chambers coined as *"greater statistics"*. There should be no conceptual barrier with regard to whether Big Data belongs to statistics or not. It could be claimed that in the near future, "big" will be the "new normal", with most kinds of data would be "big", so the term "data" would prevail again.

78. An important topic, not brought up in this paper, is the potential role of NSIs as a quality labelling authorities for statistics and indicators derived from Big Data sources. If the use of Big Data sources for the production of official statistics and indicators for public policy gains ground, it is inevitable that a body should assume some kind of a "certification" function. It is not clear whether such a function should be assumed only by NSIs or a multidisciplinary body (this would be our suggestion), but irrespectively of that, NSIs should be key members of such authorities.

[10] Within the European Commission, there are several DGs that are currently examining the potential of Big Data for their respective policy area. DG CONNECT for example are in the process of developing a Big Data platform in Horizon 2020; this initiative alone will require co-ordination and collaboration of more than 15 policy areas.